A genetic algorithm-based approach to mapping the diversity of networks sharing a given degree distribution and global clustering

Peter Overbury, Istvan Z. Kiss and Luc Berthouze

Abstract The structure of a network plays a key role in the outcome of dynamical processes operating on it. Two prevalent network descriptors are the degree distribution and the global clustering. However, when generating networks with a prescribed degree distribution and global clustering, it has been shown that changes in structural properties other than that controlled for are induced and these changes have been found to alter the outcome of spreading processes on the network. This therefore begs the question of our understanding of the potential diversity of networks sharing a given degree distribution and global clustering. As the space of all possible networks is too large to be systematically explored, a heuristic approach is needed. In our genetic algorithm-based approach, networks are encoded by their subgraph counts from a chosen family of subgraphs. Coverage of the space of possible networks is then maximised by focusing the search through optimising the diversity of counts by the Map-Elite algorithm. We provide preliminary evidence of our approach's ability to sample from the space of possible networks more widely than some state of the art methods.

Peter Overbury
Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, e-mail: po36@sussex.ac.uk

Istvan Z Kiss
Department of Mathematics, University of Sussex, Falmer, Brighton BN1 9QH, e-mail: i.z.kiss@sussex.ac.uk

Luc Berthouze
Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, e-mail: L.Berthouze@sussex.ac.uk

1 Introduction

Almost all complex systems can be modelled, to varying levels of detail, using networks whereby components of the system can be reduced down to nodes and to edges connecting them. Such an approach often makes it possible to pick out global behaviours dependent on the connections and/or relationships between different elements of the system that either would not have been noticed in isolation or could not be detected within large data sets [15]. The relationship between network structure and behaviour is the subject of much research in many areas such as epidemiology [3, 18, 9], social media [1] and neuroscience [12]. Where analytically-tractable mathematical models are needed, two main network descriptors stand out: degree distribution and global clustering. Interestingly, while there are now effective and analytically-tractable mathematical models that can handle the degree distribution well [3, 18, 9], when clustering is also considered, most models will break down or only operate for networks constructed in particular ways, e.g., networks with non-overlapping triangles [22]. This sensitivity to how networks are constructed highlights the fact that, as shown by [4, 8, 10, 19] among others, many network-generating algorithms introduce changes in structural properties other than that controlled for, thus undermining both model accuracy and inference of any causal role for the properties of interest. How to create network *null models*, i.e., where the properties of interest are fixed and all other properties are sampled in an unbiased manner, is an open question. One major step towards realising such goal would be to get a greater understanding of the space of networks satisfying a given set of requirements, e.g., a given degree distribution and a given global clustering coefficient. For networks of non-trivial size, the space of all such networks is too large to be systematically explored and therefore a heuristic approach is needed. Our approach relies on two principles: (a) a parametrisation of networks in terms of sub-graph decomposition, which significantly reduces the dimensionality of the encoding space when compared to the adjacency matrix as done in our previous work [17]; and (b) a search of the space driven by a process seeking to maximise the diversity of the networks being uncovered, thus biasing the exploration/exploitation trade-off toward exploration. The design and implementation of these two principles will be detailed in the following section.

2 Methods

2.1 Network encoding

A key challenge in exploring the space of networks satisfying constraints is that of network representation. In principle, the network's adjacency matrix would be a natural choice because it fully specifies the network. However, it suffers from two major drawbacks: scalability and unicity (two networks may have a distinct adjacency

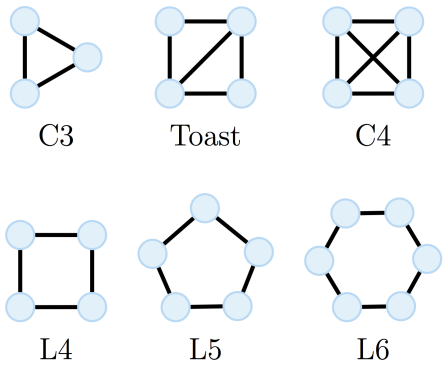

Fig. 1 The set of subgraphs used to encode networks (single edges not included). Subgraphs in the top row will induce clustering in the network.

matrix but be isomorphic). Our previous work [17] using the adjacency matrix revealed an extremely wasteful process even for small sized networks ($N = 200$). The recently-proposed dk-decomposition [16] offers an attractive alternative through its use of joint degree distributions of different orders, however, as we will show, questions remain regarding the biased nature of the network generation process once the joint degree distributions have been set. Instead, building on our recent work [20], we propose to parameterise networks in terms of a (arbitrarily chosen) family of subgraphs (see Figure 1 for a few examples).

Concretely, we use the counts of each of the subgraphs in the family to yield an adjacency matrix using the cardinality-matching algorithm (CMA hereafter) [20]. CMA is a method inspired by the configuration model [6]. It assigns a set number of subgraphs of arbitrary structure in a network with a set degree sequence. Put simply, it works by assigning to nodes in the network hyperstubs of a certain degree as specified by each subgraph in the family. For example, triangles (subgraph C3) will require 3 hyperstubs of degree 2 whereas a Toast (see Figure 1 will involve 2 hyperstubs of degree 3 (corners with 3 edges) and 2 hyperstubs of degree 2 (corners with 2 edges). These hyperstubs are then selected at random and connected until there are no more left. When a new subgraph introduces self- or multi-edges, a new node is selected as in the matching algorithm [13]. When there is no option other than to add subgraphs over existing links or selecting multiple instances of the same node, the process is restarted from scratch. To accelerate the process, in this work, only 80% of the networks' total edges were allocated to the specified subgraphs. The remaining edges were allocated as single edges to preserve the degree sequence. As this process can lead to nodes failing to have the desired degree (typically by ± 1), networks for which more than 20 nodes (out of a total of 1000) did not have the expected degree were excluded. Analysis of the networks produced (not reported here for reasons of space but available for an extended version, and see [20]) showed that the process still provides good control over most subgraphs, particularly (and advantageously in our context), those inducing clustering (i.e., C3, C4 and Toasts). Still, to avoid results being biased by a particular realisation, all measures reported in this paper were calculated by averaging over 5 network realisations. The reliability of the process is illustrated by Figure 2 which shows a compact spread of

values of three network metrics (global clustering, mean shortest path length, mean betweenness centrality) for 10,000 realisations of a single network specification.

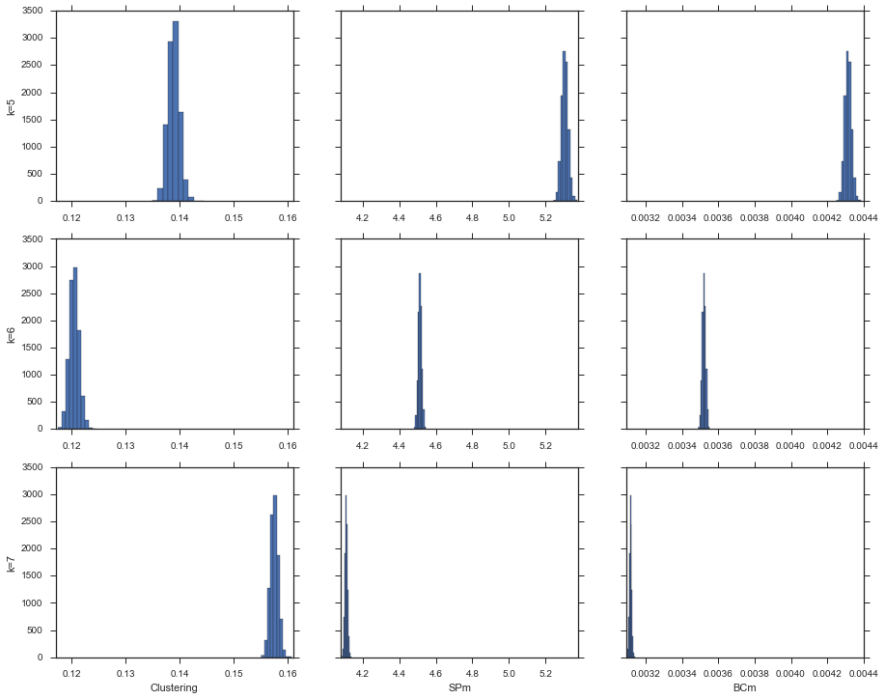

Fig. 2 Histograms of global clustering (left column), mean shortest path length (middle column) and mean betweenness centrality (right column) for 10,000 CMA realisations of a single network specification with predicted global clustering of $0.14\pm$ 0.025. The top, middle and bottom rows correspond to regular networks with degree $k = 5$, $k = 6$ and $k = 7$ respectively.

The choice of subgraphs is somewhat arbitrary and is a source of bias in itself. Here, we chose 3 subgraphs that induce clustering in the network (they are C3, C4 and toasts, see Figure 1). The other networks are loops that do not induce clustering. In this paper, only L4 and L5 were used. As a family, they provide flexibility and redundancy in the control for clustering. These 5 subgraphs have been shown in previous work to be those for which CMA showed most control over (as assessed by subgraph counting post realisation – results now shown here but available for an extended version).

2.2 Exploration of the space of possible solutions

Our primary objective being an exploration of the diversity of networks preserving a given degree distribution and global clustering coefficient, our task can be thought of as a two-part optimisation: (a) of the features that must be shared by a network for it to be added to the population of valid networks and (b) of the diversity within this population of valid networks. Multi-objective optimisation is not a new problem and the more complex variant considered here involving a changing measure of diversity within an actively changing population has recently been the focus of a number of methods in the field of genetic algorithms (GAs) [11].

In their simplest form GAs work by taking a starting population of individuals, which are encoded so that each has a *genome* that represents the key features being studied, here, the subgraph composition (expressed in percentage). This population is then *evolved* through *genetic operations* that change the genome of individuals. This typically involves *mutations* – the adding or subtracting from parts of the genome – and *recombination* or *crossover* – the combining of two individuals into a new individual with a new genome. Here, mutations involve changing the prevalence of each subgraph by a small number drawn randomly in the interval $[-0.1, 0.1]$. During crossover between two networks, a new network is created whereby a randomly chosen number of its subgraph percentages are those of the first network and all others are those from the second network. For both mutation and crossover, the subgraph prevalences of the new individual are normalised to sum up to 1. Both processes have a 60% chance of occurring to either an individual (for crossover) or an individual subgraph count (for mutation) at each generation. All individuals are then analysed for their *fitness* – the objective function in the optimisation process, here, global clustering calculated using the formula proposed in [7]. Those with the lowest fitness are either removed, selected for genetic operations less often or both. This results in a population that, depending on the setting of the GA, moves along the search space towards areas of high fitness. An important implication is that the solutions are highly dependent on the choice of the fitness measure, the selective pressures used at each generation and the way that solutions are stored.

Previous work based on the idea of optimising for diversity includes the generation of neural networks topologies for control of robots in which diversity of both behaviour and performance was optimised for [21] and our own work [17] in which we started exploring the feasibility of using GAs to optimise the diversity of networks satisfying structural constraints, albeit for small sized networks. The main limitation of these methods has been their focus on the optimisation of a few individuals to the best possible fitness over all their objectives (the Pareto front), often leading them to avoid equally valid/fit regions of the feature space. Here, we employ the recently proposed Map-Elite method [14] which seeks to map the solution space through dividing the space into identically-sized multi-dimensional *cells* that cover a set range of values for each of the features used to describe the individuals. All individuals in the population are then placed in one of these cells and when new individuals are created they are assessed based only on individuals in that same part of the space. If there is no other in the cell then the individual is deemed novel and

is kept. If, instead, there is another individual already within the cell then only the individual with the greatest fitness is kept. This method allows for the promotion of novelty without comparison of the entire population whilst also optimising the fitness of the population.

3 Results

The experiments reported in this paper sought to map the diversity of networks of size $N = 1000$ satisfying the constraint of a homogeneous/regular degree distribution (with degree 5, 6 or 7 – as three distinct scenarios) and a global clustering coefficient of 0.14. Although our choice of network encoding is insensitive to network size, the CMA connection process is not. The size $N = 1000$ makes the experiments tractable, when deployed on the high performance computing facility. The three degrees considered enable us to assess the effectiveness of the method for networks with more ($k = 7$) or less ($k = 5$) flexibility in how to allocate subgraphs. For example, with $k = 5$, it would not be possible for a node to share a fully connected square ($C5$) and the degree 3 corner of a toast whereas with $k = 7$, the same node could accommodate that and an extra free edge. Our choice of global clustering coefficient is arbitrary although one should note that depending on the choice of subgraph family used to encode networks, some clustering values are more likely than others. With the proposed family of subgraphs and the relatively small degree, it would be difficult to generate highly clustered networks, and diversity would be extremely limited. A tolerance of ± 0.025 was used in evaluating the clustering fitness of networks. A tolerance is needed due to (a) the nature of the computation of the clustering coefficient and (b) the stochasticity in allocating subgraphs and any resulting byproducts [20]. This tolerance, which is reflected in the histograms of clustering values in Figure 2, corresponds to a maximum deviation of $\pm$ 8 triangles (subgraph C3) from the expected number of subgraphs and is negligible given the number of triangles needed to achieve the required clustering.

3.1 Effectiveness of the mapping in terms of space coverage

To provide some quantitative assessment of the effectiveness of mapping, cells were configured for maximal resolution, meaning that all individuals within a cell would have the exact same subgraph counts. It should be noted at the outset (but this is currently the subject of further work) that starting out with maximal resolution is sub-optimal in terms of managing the evolutionary process. However, for the purpose of this assessment, it provides as detailed a picture as possible of the proportion of all possible encodings that is uncovered by the evolutionary process (with the caveat that with a limited number of generations, the actual number of cells uncovered can only be a tiny fraction of the total number of cells possible). In the

following, when ignoring the fact that not all combinations of subgraph counts are actually realisable – graphicality of the network), the total number of cells possible is $1040625000000 = 333 \times 250 \times 250 \times 200 \times 250$ and corresponds to the product of the ranges of possible values taken by the counts of each subgraph in the family (this count is determined on the basis of the highest-degree hyperstub in relation to the total number of nodes available in the network). The actual total number of cells is found by subtracting from the above count those cells that correspond to non-graphical/non-realisable networks, namely, those where the total number of edges prescribed by the subgraph decomposition is above $(Nk)/2$ and where the number of triple hyperstubs from C4 and Toasts is greater than $(k/3)N$ – the maximum number of triple hyper stubs allowed by CMA in a network. Coverage of the space at various points during the process is shown in Table 1. Given the maximum resolution and the fact that each generation only produces one new network, the actual percentage of coverage is very small. However, the table shows two important results: (a) the rate at which new cells are explored in relation to the number of generations is almost 1 suggesting that cells are not revisited (this would no longer be the case if cells had lower resolution); (b) the rate at which valid networks are produced is roughly constant as the number of generations increases.

k	21,000 gen		42,000 gen		63,000 gen	
	Explored	Valid	Explored	Valid	Explored	Valid
5	20783	12995	41546	25952	62286	38852
6	20824	18266	41583	36596	62349	55009
7	20845	18691	40646	36680	62431	56435

Table 1 Number of explored and valid cells uncovered by the evolutionary process at various time points for the three scenarios ($k = 5, 6, 7$) considered. In all cases, networks have size $N = 1000$ and the family of subgraph considered is (C3, C4, Toast, L4 and L5) with a desired global clustering of 0.14 ± 0.025. For reference, the total number of cells possible (after removal of non-graphical solutions) is $\sim 10^{12}$. Each generation can produce at most one new network.

Importantly, we note that this table does not provide any information regarding coverage of the space of valid networks, those with correct degree distribution and global clustering within ± 0.025 of the desired clustering. Whilst the search is focused on finding valid cells (rather than all possible cells), we do not have any estimate for the total number of possible valid networks in the space of all possible networks. Figure 3 provides a different perspective on this by using low-dimensional projections of the space of networks explored and valid. Where possible, non-graphical solutions have been highlighted. The Figure reveals that despite the limited number of generations (again, corresponding to a very small percentage of all possible configurations) there is evidence of fairly uniform sampling as far as explored cells are concerned. The Figure further reveals pair-wise relationships between counts of subgraphs that reflect the constraints of the problem. For example, when two clustering-inducing subgraphs are considered (e.g., C4 and Toast) there is a distinct relationship whereby configurations with larger numbers of C4s have smaller numbers of Toast and conversely. Instead when clustering-inducing

subgraphs and non clustering-inducing subgraphs are considered (e.g., C3 and L4) valid configurations can be found throughout the space of explored solutions. Areas that are not explored are typically reflecting configurations for which although no graphicality condition is being violated as far as the particular pair of subgraphs is concerned, no network realisation is possible when taking into account the other dimensions.

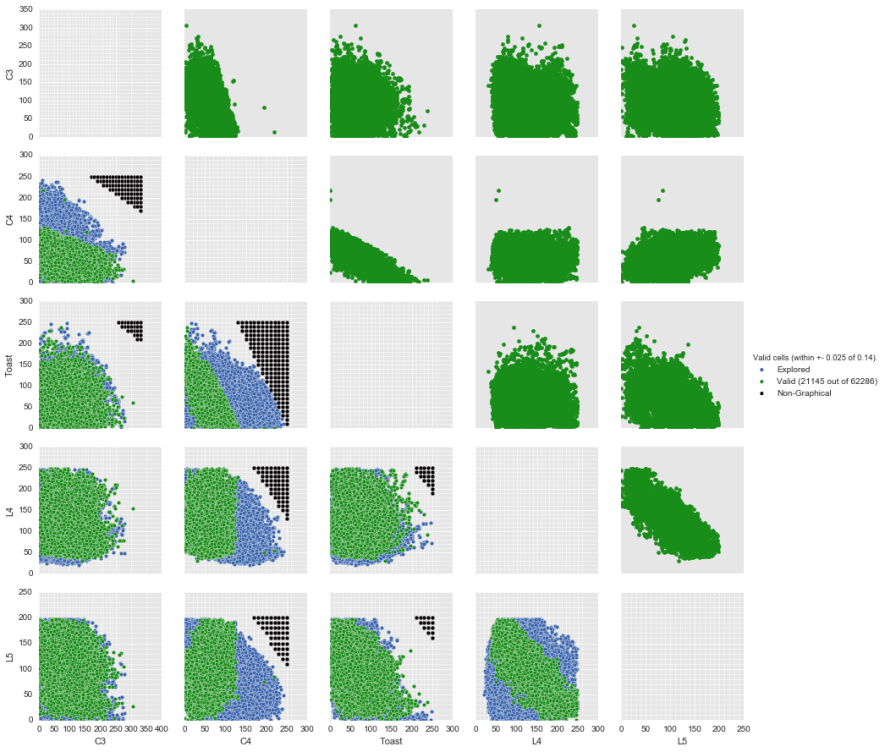

Fig. 3 Low-dimensional projections of the configurations discovered by the evolutionary process (both those that were explored but not necessarily satisfying the constraints – in blue – and those that were valid – in green) after 63040 generations. Each dot denotes a network whose coordinates are the counts for the subgraphs shown in the horizontal and vertical axes. A dot does not define a unique network, however, as the projection can mask great diversity in the remaining 3 dimensions.

3.2 Comparison with other methods

Whilst the above results point to evidence of diversity in terms of subgraphs a more useful basis for evaluating the effectiveness of our approach is to assess the extent to which networks uncovered show greater diversity than can be expected from meth-

ods currently available to generate networks satisfying the same constraints. Since subgraphs counts are explicitly controlled by the evolutionary process, they would not be a fair metric for comparison. Instead, we considered two global structural properties: mean shortest path length and mean betweenness centrality (BCm) – although as both show a high degree of correlation, only betweenness centrality will be reported below. These properties are important determinants of behaviour in networks [15]. Two state of the art network generating methods have been used for this comparison: dk-series decomposition [16] and BigV rewiring [5]. For the former, we used dk2.1 (using code from [2]) which preserves degree distribution and global clustering (dk2.5 would also preserve local clustering which is overly specific for our purpose). Since the dk method requires a seed network to operate, one network was chosen at random among those generated by our approach. For the latter, the rewiring algorithm was applied to a single random network with homogeneous degree distribution who was rewired until desired clustering was achieved (with a maximum of 40000 rewirings). For both BigV rewiring and dk decomposition, the number of networks generated was set to the number of networks produced by the GA.

Figure 4 reveals that the range of mean betweenness centrality for networks produced by our approach is greater than that of either (or even both of) the dk- and BigV-produced networks, suggesting that a wider area of the space of solutions was explored. This holds for all three scenarios ($k = 5, 6, 7$). An important correlate of this finding is that neither BigV rewiring nor dk-decomposition can claim to generate null models. Interestingly, the networks produced by both methods do not appear to overlap suggesting that either methods generate networks in different areas of the space of solutions. Likewise, although our method appears to sample more widely than BigV rewiring and dk, full overlap only occurs for $k = 7$ whereas there is almost no overlap for $k = 5$. It remains to be seen whether, given more time, our method would uncover these areas of the space of solutions. Finally, given that the dk networks were produced from a single seed, it is worth pointing out that there was no obvious correlation between the betweenness centrality of the seed and the mean betweenness centrality for the dk-generated networks. The extent to which the choice of seed conditions the distribution of networks generated remains unclear.

4 Discussion

In this paper, we have proposed a new GA-based approach to generating networks preserving degree distribution and global clustering. Our approach is focused on maximising the diversity of the networks being created. Since it is impossible to quantify the extent to which the entire space of solutions has been sampled, we have provided evidence of the effectiveness of the method by comparing it to two state of the art network-generating methods, dk-series decomposition and BigV rewiring and showing that our method generates more diversity. Whereas coverage of the space of solutions using our method will depend on the number of generations avail-

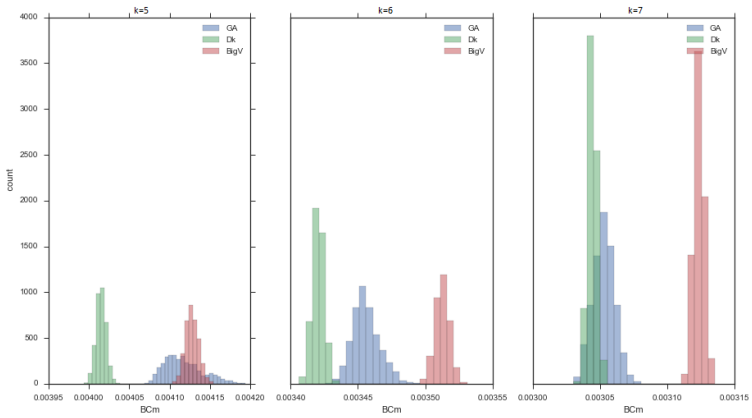

Fig. 4 Histograms of the mean betweenness centrality for the proposed method (blue), BigV rewiring (red) and dk2.1 (green) for each of the three scenarios: $k = 5$ (left), $k = 6$ (middle), $k = 7$ (right). The same number of networks was used for all three methods.

able, both BigV rewiring and dk-series decomposition depend on a mixing time being reached. Care must therefore be taken in making definite statements about the ability of these methods to sample the range of networks found by our approach. However, given the same number of steps, there was greater diversity using our approach. This provides evidence for the usefulness of our method in the evaluation of the level of bias shown by current network generation methods. Much further work is needed to strengthen our framework, especially given that it is itself subject to a number of biases. For example, whilst encoding in terms of subgraphs provides much flexibility and scalability, it is itself a source of biases. At this time, it is unclear how a different choice of family would affect the diversity of networks uncovered. On the bright side, we believe that our starting scenario of networks with homogeneous distribution and low degree actually made it much harder to find diversity in the networks. The immediate focus will be to consider heterogeneous distributions with higher degrees. Whilst it will not affect computation time, it will provide much more flexibility for the network connection process (CMA) to realise networks (as well as remove the need to allow for 20% free edges, thus providing further control).

References

1. Aggarwal, C. C. (2011). An introduction to social network data analytics. In Social network data analytics (pp. 1-15). Springer US.
2. Colomer de Simón, P. (2014). RandNetGen [Computer software]. Retrieved from https://polcolomer.github.io/RandNetGen/. Last accessed 14 September 2016.

3. Danon, L., Ford, A.P., House, T., Jewell, C.P., Keeling, M.J., Roberts, G.O., & Vernon, M.C. (2011). Networks and the epidemiology of infectious disease. Interdisciplinary perspectives on infectious diseases, 2011.
4. Green, D. M., & Kiss, I. Z. (2010). Large-scale properties of clustered networks: Implications for disease dynamics. J Biol Dyn, 4(5), 431-445.
5. House, T., & Keeling, M.J. (2010). The impact of contact tracing in clustered populations. PLoS Comput Biol, 6(3), e1000721.
6. Karrer, B., & Newman, M.E. (2010). Random graphs containing arbitrary distributions of subgraphs. Phys Rev E, 82(6), 066118.
7. Keeling, M.J. (1999). The effects of local spatial structure on epidemiological invasions. Proc R Soc Lond B: Biol Sci 266(1421), 859867.
8. Kim, H., Toroczkai, Z., Erds, P.L., Mikls, I., & Szkely, L.A. (2009). Degree-based graph construction. J Phys A-Math Theor, 42(39), 392001.
9. Kiss, I.Z., Miller, J.C., & Simon, P.L. (in Press). Mathematics of epidemics on networks: From exact to approximate models, Springer.
10. Klein-Hennig, H., & Hartmann, A. K. (2012). Bias in generation of random graphs. Phys Rev E, 85(2), 026101.
11. Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. Evol Comput, 19(2), 189-223.
12. Mears, D., & Pollard, H. B. (2016). Network science and the human brain: Using graph theory to understand the brain and one of its hubs, the amygdala, in health and disease. J Neurosci Res, 94(6), 590-605.
13. Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.E., & Alon, U. (2003). On the uniform generation of random graphs with prescribed degree sequences. arXiv preprint cond-mat/0312028.
14. Mouret, J.B., and Clune J. (2015). Illuminating search spaces by mapping elites. arXiv preprint arXiv:1504.04909.
15. Newman, M.E.J. Networks: An introduction. Oxford University Press, 2010.
16. Orsini, C., Dankulov, M.M., Colomer-de-Simón, P., Jamakovic, A., Mahadevan, P., Vahdat, A. & Fortunato, S. (2015). Quantifying randomness in real networks. Nat Comm, 6:8627.
17. Overbury, P., & Berthouze, L. (2015, July). Using novelty-biased GA to sample diversity in graphs satisfying constraints. In Proceedings of the Companion Publication of the 2015 Annual Conference on Lect Notes Comput Sc (pp. 1445-1446). ACM.
18. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. Rev Mod Phys, 87(3), 925.
19. Ritchie, M. and Berthouze, L. and Kiss, I.Z. (2016). Beyond clustering: Mean-field dynamics on networks with arbitrary subgraph composition. J Math Biol 72(1-2), 255-281.
20. Ritchie, M. and Berthouze, L. and Kiss, I.Z. (2016). Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters. J Complex Networks, cnw011.
21. Stanley, K.O., & Miikkulainen, R. (2003). A taxonomy for artificial embryogeny. Artif Life, 9(2), 93-130.
22. Volz, E.M., Miller, J.C., Galvani, A., & Meyers, L.A. (2011). Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput Biol, 7(6), e1002042.